Dependency parsing for interaction detection in pharmacogenomics

Schneider, Gerold ; Rinaldi, Fabio ; Clematide, Simon

Abstract: We give an overview of our approach to the extraction of interactions between pharmacogenomic entities like drugs, genes and diseases and suggest classes of interaction types driven by data from PharmGKB and partly following the top level ontology WordNet and biomedical types from BioNLP. Our text mining approach to the extraction of interactions is based on syntactic analysis. We use syntactic analyses to explore domain events and to suggest a set of interaction labels for the pharmacogenomics domain.

Originally published at:
Schneider, Gerold; Rinaldi, Fabio; Clematide, Simon (2012). Dependency parsing for interaction detection in pharmacogenomics. In: LREC 2012: The eighth international conference on Language Resources and Evaluation, Istanbul, 21 May 2012 - 25 May 2012.

Dependency parsing for interaction detection in pharmacogenomics

Gerold Schneider, Fabio Rinaldi, Simon Clematide

Institute of Computational Linguistics, University of Zurich,
Binzmühlestrasse 14, 8050 Zürich, Switzerland
{gschneid,rinaldi,siclemat}@cl.uzh.ch

Abstract

We give an overview of our approach to the extraction of interactions between pharmacogenomic entities like drugs, genes and diseases and suggest classes of interaction types driven by data from PharmGKB and partly following the top level ontology WordNet and biomedical types from BioNLP. Our text mining approach to the extraction of interactions is based on syntactic analysis. We use syntactic analyses to explore domain events and to suggest a set of interaction labels for the pharmacogenomics domain.

Keywords: Pharmacogenomics, Event Classes, Interaction Detection

1. Introduction

Pharmacogenomics is the discipline which studies the mutual interactions among drugs, genes, diseases, in particular in relation to specific individual mutations, which can affect the reactions to drugs and the susceptibility to diseases. One important database that aims at providing a reference repository for such information is PharmGKB (Sangkuhl et al., 2008). The information contained in PharmGKB is obtained from a combination of submitted experimental results and literature curation. Literature curation is the knowledge-intensive process which aims at extracting from the primary literature (scientific publications) the most relevant results obtained by the authors in their scientific experiments. Despite some support by text mining tools, it is still the case that the process of curation involves extensive human intervention, which is time consuming and expensive.

In this paper we describe research conducted by the OntoGene group within the scope of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature[1]), which aims at producing novel efficient text mining tools which provide better support for the process of biomedical literature curation. In particular we have recently used the PharmGKB database in order to derive interaction indicators from the literature. The OntoGene research group has participated in several text mining shared tasks, such as BioCreative (Rinaldi et al., 2008; Rinaldi et al., 2010b; Schneider et al., 2011), CALBC (Rinaldi et al., 2010a) and BioNLP (Kaljurand et al., 2009), which present structural similarities with the extraction of interactions in the pharmacogenomics domain, as we discuss later.

We describe applications of the text mining technologies developed for the problem of finding head words (so called "triggers") and categorize entity interactions into classes relevant to the pharmacogenomics domain.

2. The OntoGene text mining system

Biomedical researchers studying various biological processes need to find supporting evidence for specific relationships among entities of interest, such as protein-protein interactions, or influence of genes on specific diseases. These activities can profit from text mining systems, which not only can find relevant publications, but also deliver small passages describing the interactions that they need. This capability obviates the need to read entire documents, and allows researchers to find answers to their questions more quickly. Many interaction detection approaches in the pharmacogenomics domain use untyped interactions, or the labels mirror the types of the participants such as drug or disease. Often, interactions could be classed into meaningful types. For example, proteins and genes, may bind, block, inhibit etc. This need is recognized by some of the biomedical text mining competitions, for example BioNLP (Cohen et al., 2009), which uses classes of interactions, and finding the interaction class label is an integral part of the competition.

Intuitively, interactions between other biomedical entities also fall into clearly distinguishable classes. For example, a certain gene can increase the risk for a disease, a certain drug can inhibit a gene, or have a healing effect on a disease, or have side-effects. Most approaches going beyond gene and protein interactions use unlabelled interactions, or at best the interaction type follows deterministically from the involved entitities. We believe that a more detailed inventory of classes is beneficial and feasible.

2.1. Our syntax-based approach

Approaches towards identification of entity interactions based on their coccurrence in a given text span are quite common (e.g. (Rebholz-Schuhmann et al., 2006)). Some approaches apply handcrafted rules, for example regular expressions for surface searches (Giuliano et al., 2006), or syntactic patterns on automatically parsed corpora (Rinaldi et al., 2006; Fundel et al., 2007a). These approaches typically achieve high precision at the cost of recall. There have recently been numerous publications showing the potential of dependency-based language analysis for text mining (e.g. (Clegg and Shepherd, 2007; Fundel et al., 2007b)). (Pyysalo et al., 2007) describes a manually annotated corpus which includes a dependency based analysis of each sentence. (Clegg and Shepherd, 2007) uses dependency graphs in order to benchmark four publicly available

[1] http://www.sasebio.org/

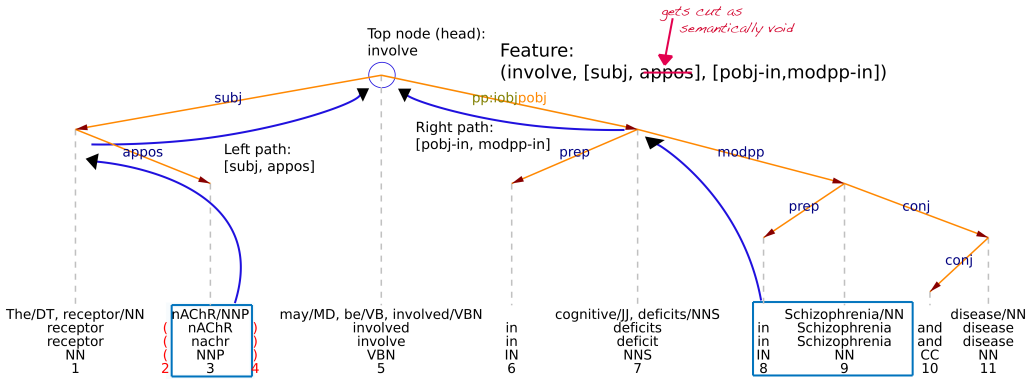

Figure 1: Simplified internal syntactic representation of the sentence "The neuronal nicotinic acetylcholine receptor alpha7 (nAChR alpha7) may be involved in cognitive deficits in Schizophrenia and Alzheimer's disease." from PubMed abstract 15695160. The curved arrows and dark red notes are aimed at illustrating the path features.

natural-language parsers. (Fundel et al., 2007b) describes a large-scale relation mining application using the Stanford Lexicalized Parser.

These approaches can be further enhanced using machine learning methods, by extracting meaningful features from the dependency parse trees and from other intermediate stages of processing (e.g. (Erkan et al., 2007; Kim et al., 2008; Van Landeghem et al., 2008)).

In previous work, we used manually-constructed syntactic patterns in order to filter candidate protein-protein interactions (Rinaldi et al., 2007; Rinaldi et al., 2008). This approach was later enhanced with automatic learning of useful syntactic configuration from a training corpus (Schneider et al., 2009; Rinaldi et al., 2010b). In the following we describe how such an approach has been adapted to PharmGKB.

We have parsed all sentences in the PharmGKB gold standard with our own dependency parser (Schneider, 2008). Entities are recognized and disambiguated using the OntoGene pipeline. All entities that appear in the same sentence are potentially interacting, so we record the syntactic path that connects them as *candidate path*. If the gold standard contains the information that these two entities really interact, then we mark the path that connects them as *relevant path*. The calculation of candidate path divided by relevant path gives us the Maximum-Likelihood probability that a path is relevant:

$p(relevant|candidate\ path) = \frac{f(relevant\ path)}{f(candidate\ path)}$

The most frequent path types in the training set are given in table 2. The third line, where the head word is *effect*, for example, has a modification by an of-PP to one of the entities in the relation, and a nested on-PP and of-PP modification. It covers patterns like *the effect of X on the increase of Y* or *no effect of X on the development of Y*, where X and Y are domain entities like drug, disease and protein. The most frequent dependency types of the Pro3Gres parser are given in table 1. The dependency set is closely related to GREVAL (Carroll et al., 2003), to which we have mapped for evalutations (Schneider, 2008). It can also be mapped to the Stanford scheme (Haverinen et al., 2008).

The first column of table 2 contains the probability $p(relevant|candidate\ path)$. We can use this probability directly during the application phase: whenever two entities occurring in the same sentence of the application corpus, for example a drug and a disease, have a probability of being relevant above a certain threshold, the systems reports the interaction. p(relevant) can then be interpreted as the potential precision of such direct application. As syntactic path, we record the dependency labels that connect the two entities, and the topmost word connecting them. A sample path is provided in Figure 1.

RELATION	LABEL	EXAMPLE
verb–subject	*subj*	*he sleeps*
verb–direct object	*obj*	*sees it*
verb–second object	*obj2*	*gave (her) kisses*
verb–adjunct	*adj*	*ate yesterday*
verb–subord. clause	*sentobj*	*saw (they) came*
verb–pred. adjective	*predadj*	*is ready*
verb–prep. phrase	*pobj*	*slept in bed*
noun–prep. phrase	*modpp*	*draft of paper*
noun–participle	*modpart*	*report written*
verb–complementizer	*compl*	*to eat apples*
noun–preposition	*prep*	*to the house*

Table 1: Frequent Pro3Gres dependency types

Such a direct application, however, suffers from sparse data problems. If possible, we use a single feature for the entire path. In the majority of cases, we need to split the path into two halves: from the top-word down to one of the entities as feature 1, and from the top-word down the other entity as feature 2.

We also use lexical information on transparent words (Meyers et al., 1998) to avoid data sparseness, as follows:

- First, entities occurring inside noun chunks are allowed to replace the head of the chunk if the head is a transparent word.
- Second (if still no relevant path exists), the relations for appositions, conjunctions and hyphens are cut.
- Third (if still no relevant path exists), parts of trees which are headed by a transparent word are cut.

p(relevant)	Head	Path1	Path2	TP	Count
13.62%	associate	subj	pobj-with	53	389
17.82%	associate	subj modpp-in	pobj-with	31	174
18.92%	effect	modpp-of	modpp-on modpp-of	21	111
20.65%	association	modpp-of	modpp-with	19	92
6.29%	be	obj modpp-of	subj	19	302
17.82%	metabolize	pobj-by	subj	18	101
29.63%	inhibit	pobj-by	subj	16	54
23.81%	cause	subj modpp-in	obj	15	63
100.00%	analyze	subj modpp-in	pobj-in modpart pobj-with	14	14

Table 2: Some of the most frequent path types in the training set

A transparent word is a word that does not affect the meaning of sentence fundamentally if it is left out. For example, if *drug A affects groups of patients* then the sentence *drug A affects patients*, which does not contain the transparent word *group*, has a very similar meaning. We have learnt transparent words using a machine learning approach: words that occur particularly often inside paths are regarded as transparent (Schneider et al., 2009).

Three additional factors are used to calculate a score. First $f(c_1)$, the frequency of the entities in the document, as the most relevant entities in the given document are typically mentioned several times. Reporting interactions based on the frequencies of entities leads to a very high baseline in protein-protein interaction (Rinaldi et al., 2010b). Second $f(c_2)$, the probability of the entity types to enter interactions is used. For example, the probability that a drug and a disease in the same sentence have an interaction is relatively high (about 12%), while the probability that two drugs appearing in the same sentence interact is low (about 1%). Third, we use a simple zoning factor: the title is given ten time the weight of the rest of the text.

A score is assigned to every candidate interaction according to the following formula:

$pscore(c_1, c_2) = p(relevant|candidate\ path) * f(c_1) * f(c_2) * p(relevant|entity\ types) * zoningfactor$

This score is based on probabilites, but it does not express a probability. It can be used for ranking candidates, report those above a threshold and use it as confidence measure, for example for ranking different interactions that may be expressed in the sentence. Our syntax-based approach in its current version only has two backoffs: it splits the path into a left and right half, and transparent words are filtered. It can reach relatively high precision at the cost of low recall. The path contains important information on the type of interaction as we discuss in section 3.

2.2. Evaluation

We have applied our approach to a manually verified test set from the pharmacogenomics domain. In collaboration with PharmGKB we conducted a separate experiment to test the usefulness of our text mining technologies and curation interface for a simple revalidation experiment which is described in detail in (Rinaldi et al., 2012). This experiment produced abstracts where all interactions have been reliably curated by PharmGKB domain experts. We have used 75 of these documents as a test corpus, and the rest of the PharmGKB dataset was used for training purposes, excluding also all documents that contain more than 20 interactions. Evaluation results are given in table 3.

The method **syn** is purely our syntactic method, as described in section 2.1. We see that it has higher precision than recall. Recall can be increased by including sentence-coocurrence, which the method **syn+cooc** does. We can see on the one hand that recall increases at the cost of precision, on the other hand that it is still below 50%, which indicates that many interactions are expressed across several sentences. The method **syn+cooc2** extends the sentence-coocurrence score to including the neighbouring sentence. The increase in recall indicates that context of more than one sentence is often necessary. The method **syn+cooc2w** weighs the sentence-coocurrence score by distance, giving higher scores to entities that appear closer. The method **syn+cooc2wf** is identical but does not use a score threshold, thus returning all results, which increases recall and reduces precision. It aims to give an upper bound on recall. Results using only the first n reported hits are also given. The method **syn+cooc2wb** is identical to **syn+cooc2wf** but uses a relatively high score threshold aiming for a precision/recall balanced output.

In addition to being a useful component of an interaction detection approach, the syntactic approach detects the lexemes appearing at the top of the syntactic path, as we discuss now.

3. Classifying drug-gene-disease interactions

3.1. Data-driven exploration of trigger words

The most frequent true positive types are given in table 2, broken down by left-path, right-path and top word, i.e. the word at which the two paths from the entities up to the root node meet. The top word is often the keyword expressing that an interaction takes place, the so-called trigger word. The counts are sorted by inverse frequency, the most frequent path type has 53 instances. Path1 is the half from the top word (Head) of the path to the first entity, path2 the half to the second entity. The last column lists how often the path occurs in the entire training corpus, irrespective whether it expresses relevant interactions or not, which we refer to as candidate path. The probability $p(relevant|candidate\ path)$, which is the main factor in the syntactic feature, is given in the first column. We can

Method	Docs	TP	FP	FN	AUCiP/R	n	P	R
syn	43	36	149	116	0.215	all	0.307	0.286
syn+cooc	73	116	1044	151	0.277	all	0.143	0.477
syn+cooc2	72	158	2337	106	0.279	all	0.094	0.616
syn+cooc2w	72	165	2685	99	0.286	all	0.091	0.650
syn+cooc2wf	72	23	49	241	0.103	1	0.319	0.103
syn+cooc2wf	72	37	107	227	0.154	2	0.257	0.170
syn+cooc2wf	72	45	171	219	0.175	3	0.208	0.205
syn+cooc2wf	72	67	293	197	0.215	5	0.186	0.312
syn+cooc2wf	72	101	611	163	0.257	10	0.143	0.444
syn+cooc2wf	72	167	3783	97	0.286	all	0.073	0.661
syn+cooc2wb	53	47	180	147	0.220	all	0.270	0.281

Table 3: Evaluation of 75 manually annotated documents. The first column gives the approach used. The second column reports the number of documents with a least one response hit. The third to the fifth column give true positives (TP), false positives (FP) and false negatives (FN). The sixth column contains the macro averaged AUCiP/R. The seventh column contains the cut-off value n used by the BioCreative evaluation tool as a threshold on the number of response hits when computing these results. In rows with $n = all$ no threshold was applied. The eighth column reports macro precision, the ninth macro recall.

p(relevant)	Head	Path1	Path2	TP	Count
100.00%	analyze	subj modpp-in	pobj-in modpart pobj-with	14	14
100.00%	investigate	subj modpp-of	sentobj obj modpp-with modpp-of	12	12
100.00%	effect	bridge modpp-of	modpp-on modpp-of	6	6
100.00%	determine	bridge	subj nchunk modpp-for modpp-of	5	5
100.00%	involve	subj	pobj-in modpp-in	4	4
90.00%	disease	nchunk	chunk(genes)	9	10
88.89%	explain	subj	pobj-in	8	9
83.33%	determine	bridge	sentobj subj	5	6
83.33%	catalys	subj	bridge obj	5	6
83.33%	cancer	modpp-in	chunk(risk)	5	6
80.00%	effect	modpp-of	bridge modpp-on modpp-of	4	5
66.67%	metabolise	subj	bridge	4	6
66.67%	measure	sentobj subj modpp-of	bridge	4	6
66.67%	find	obj modpp-between	obj2 modpp-with	4	6
66.67%	determine	subj modpp-in	obj modpp-in modpp-to	4	6
66.67%	correlate	pobj-in	subj	4	6
66.67%	be	pobj-in	obj modpp-of modpp-between	4	6
60.00%	investigate	bridge modpp-of	obj modpp-of	6	10

Table 4: Syntactic paths with high probability of expressing an interaction

see, for example, that the verb *be* is generally unlikely to head a relevant path, while *cause*, *association*, *associate*, and *analyze* have much higher probabilities. Also obvious, short and easily interpretable paths such as the first one of table 2 (*"X associates with Y"*) only have relatively low chances of expressing relevant entities, which indicates that naive implementations of the syntactic feature would have low precision. The very specific and long path in the last row always expresses a relevant interaction. There are 15 paths occurring more than 3 times which have a 100% probability.

A benefit of the syntactic approach is that it detects the lexemes appearing at the top of the path (column 'Head' in the tables), which can be used as keywords for other approaches and may also help to distinguish interaction classes. All paths that are not cases of self-reference and are relevant with at least 60% are given in table 4.[2] Except for *be* in a very specific configuration, all head words in table 4 are good keyword candidates. In the case of *be* the words inside the path often contain interaction type information. In table 5 we see, for example, that there are 30 cases in which a drug *is* an inhibitor of a gene. As classes for gene and protein interactions have already been suggested, we restricted the interactions in table specifically to drugs, genes, and/or diseases.

Figure 1 portrays a gold standard interaction which correspond to the fifth row in table 4. The gene-disease interaction between *'nAChR'* and *'Schizophrenia'* (and also *'Alzheimer's disease'*) is expressed in this sentence. Path1

[2]Two dependency types appearing in this table were not explained in table 1. The type *nchunk* repairs underchunking, the type *bridge* connects partial parses.

Count	entity1	words1	entity2	words2
45	DISEASE	associate	GENE	associate
30	DRUG	inhibit	GENE	inhibit
28	DISEASE	association	GENE	association
27	DRUG	effect	GENE	effect
24	DRUG	be	GENE	be
23	DRUG	influence	GENE	influence
23	DRUG	associate	GENE	associate
23	DISEASE	associate	GENE	associate polymorphism
19	DISEASE	cause	GENE	cause mutation
16	DISEASE	cause	GENE	cause
14	DRUG	transport	GENE	transport
14	DISEASE	associate risk	GENE	associate
12	DRUG	investigate follow treatment modulator	GENE	investigate expression
12	DRUG	be	GENE	be inhibitor
10	DISEASE	treat	DRUG	treat

Table 5: Words in syntactic paths connecting entity types (selected examples).

leads via apposition and subject relation to the verb 'involve'. The apposition relation is semantically void and thus gets cut. Path2 is up from 'Schizophrenia' via the relations modpp-in and pobj-in to 'involve', which is suggested as the head because the paths meet here. Head words like 'involve' are quite unspecific, the type of interaction is left underspecified. The verb group (*may be involved*) clarifies, however, that the class of interaction is the subject of the investigation. A possible interaction could thus be 'speculation'. The top rows in table 2 (*associate, effect, association*) are also unspecific. Looking at the data, however, reveals that the article context specifies the role in the vast majority of cases, although often outside the clause containing the interaction type. The first three instances of the top row, for example, are:

(1) " **In conclusion, our data suggest** *that the TT MTHFR 677 genotype is associated with marked MTX - induced hyperhomocysteinemia ...* ";

(2) " *In cell-based , transactivation assays , OATP-C expression was associated with* **increased** *cellular rifampin retention ...* ";

(3) " **The objective of this study was to evaluate whether** *the MDR1 exon21 and exon26 polymorphisms and the CYP3A5 polymorphism are associated with tacrolimus disposition ...* ";

Markers pointing to specific interpretations are given in boldface. The sentences indicate that specific interpretations such as 'increase' (example 2), 'decrease' or 'speculation' (examples 1 and 3) are often intended, but the task of detecting them can be demanding. In cases where no specific interpretation marker exists, the default for associate is usually 'increase'. Looking again at the data, the first instance of the second row of table 2 illustrates this.

(4) " *A coding polymorphism in the receptor with* **reduced** *affinity to LTD4 is associated with asthma.* ";

The fact that the association with asthma is positive is not specified, but the negative affinity (BioNLP class *binding*) is explicitly marked. 'increase' marks a positive association, 'decrease' a negative association.

Sometimes, it is explicitly underspecified whether an association is positive or negative, as in the following example. The polarity of the association is not mentioned, the polarity of the expression is explicitly underspecified.

(5) " *Some genetic studies have found that C - to-T single-nucleotide polymorphism (C-509T) in the TGF-beta1 gene promoter may be associated with* **altered** *gene expression and asthma phenotype .* ";

Some of the specific roles directly appear as the head word, for example *inhibit, cause, increase, treat, risk* in table 5. *Cause* and *increase* can be seen as positive association, *inhibit* as negative association, and risk as speculative association. A head word like 'treat' could even be seen as speculative positive association (because it is hoped that the patient's health will improve), but it is ontologically difficult to assess how far one can subsume events under the same label.

3.2. Event Ontologies

Research on semantic primitives (Wierzbicka, 1996) is often contested (Goddard, 1998) but as a coarse-grained operationalization for IE purposes they are certainly useful. For example the top-level ontology WordNet (Miller et al., 1990) distinguishes 15 lexicographer files which form lexical verb classes. They are given in table 6. In the last column we give relevant examples from the pharmacogenomics domain.

WordNet is a top level ontology. For the pharmacogenomical domain, many of the WordNet ontology classes are not used, they are only partly useful for a domain ontology. The relevant events cluster in classes 29, 30, 31 and 41.

File 29 and 30 partly overlap, as we have discussed: *heal* appears in both classes; and *treat* can be a speculative positive association, and might then also fall into class 30.

File 30 is too general, containing research methods (*process*), positive and negative associations (*increase, heal, decrease*) and chemical processes (*bind, transcribe*). Biomedical and chemical processes should stay more fine-grained in the pharmacogenomics domain, for example following the BioNLP event classes. Many biomedical events,

File Number	Name	Contents	Examples from the pharmacogenomicsl domain
29	verb.body	verbs of grooming, dressing and bodily care	treat, heal (get healthy again), recover, cure
30	verb.change	verbs of size, temperature change, intensifying, etc.	process, increase, decrease, heal (mend), recover, bind, transcribe
31	verb.cognition	verbs of thinking, judging, analyzing, doubting	analyze, examine, study, associate, prove, show
32	verb.communication	verbs of telling, asking, ordering, singing	investigate
33	verb.competition	verbs of fighting, athletic activities	
34	verb.consumption	verbs of eating and drinking	
35	verb.contact	verbs of touching, hitting, tying, digging	
36	verb.creation	verbs of sewing, baking, painting, performing	cause
37	verb.emotion	verbs of feeling	
38	verb.motion	verbs of walking, flying, swimming	
39	verb.perception	verbs of seeing, hearing, feeling	
40	verb.possession	verbs of buying, selling, owning	
41	verb.social	verbs of political and social activities and events	associate, risk
42	verb.stative	verbs of being, having, spatial relations	
43	verb.weather	verbs of raining, snowing, thawing, thundering	

Table 6: WordNet Lexicographer file classes for verbs, with pharmacogenomic event verb examples in the last column.

for example *express* and *localize* are not present in WordNet in their biomedical sense.

File 36 contains verbs like *cause* and *make*, and is also an important concept in the biomedical domain. The senses given for *associate* in WordNet are on the one hand mental connections in class 31, or social activities *"He associates with strange people"* in class 41. Examples 1-5 all use a reading of "associate" in the sense of correlation, which expresses *increase* or *decrease*, and actually falls into file 30. While *risk* may refer to dangerous social behaviour also in the pharmacogenomics domain, it can equally be meant as potential, speculative causing factor. Genetic factors, for example, are a risk that is not caused by social behaviour, and could also be subsumed to file 36.

To summarize, based on our inspection of the paths in tables 3 and 4, the example sentences that they cover, and additional random samples further down in the lists we find that the interaction classes that have been suggested for the interaction of genes and proteins, for example in the BioNLP shared task, are useful, but they do not cover all cases that are needed in the broader pharmacogenomics domain. The top level ontology, on the other hand, is too broad; it suffices to use a subset of the available classes. We would like to suggest the following additional event labels for the pharmacogenomics domain:

- *treat*: WordNet file 29. Could also be conflated with the following event class
- *associate/increase/decrease/correlate*: as used in examples (1) and (2), WordNet file 30
- *analyze/examine/investigate/show/prove*: as used in (3), WordNet file 31
- *cause/risk*: WordNet file 36

3.3. Polarity and speculation

We have mentioned that WordNet file 30 and our suggested *associate* class is extremely broad. *Increase* or *heal* express positive associations, *decrease* a negative association. They refer to the same type of event, but with opposite polarity.

We have suggested to keep the BioNLP event type labels. They are:

- Gene Expression
- Transcription
- Protein Catabolism
- Phosphorilation
- Localization
- Binding
- Regulation

The BioNLP label *Regulation* has the additional types *Positive Regulation* (for example activation) and *Negative Regulation* (for example inhibition). They can be seen as subclass of the *Regulation* type and express the inherent polarity of an event. Research on detecting the inherent polarity of events has been popular recently under the name of sentiment detection. Inherent polarity is orthogonal to the event type, we would therefore suggest to use orthogonal additional classes, and only use one *Regulation* class.

BioNLP uses orthogonal classes for expressing polarity in the form of negation and speculation. There has recently been increased interest in detecting negation and speculation in the biomedical domain, for example (Sarafraz and Nenadic, 2010).

Both polarity and speculation can be expressed explicitly or be inherent properties of an event. Negative polarity is explicitly expressed by a negation, speculative polarity by using modal verbs such as *may* (example (5)) and *could*. Inherent negative polarity is e.g. expressed by *decrease* for *associate*, inherent speculative polarity by word semantics, for example *we show* versus *we investigate* or *cause* versus *risk*. Often polarity is supported by the syntactic context (*we investigate* **whether** or **In conclusion**, *we show*). We suggest these classes as a working hypothesis. Since we have derived them from frequent syntactic patterns and from clear textual features, they could strike a reasonable balance between too specific and too general, and they could probably be detected by text mining approaches.

4. Conclusion

Based on our pilot study, we believe that interaction classes are both beneficial and feasible. The set of classes will need to be discussed and tested in more detail in future research. We have suggested the following labels as a working hypothesis: *treat* (WordNet file 29), *associate/correlate* (WordNet file 30), *investigate* (WordNet file 31) and *cause* (WordNet file 36) in addition to the BioNLP labels. The orthogonal modification labels *speculation* and *polarity* have also been suggested to play an important role, and allowing

us to reduce the number of event classes that are needed. *risk* is e.g. a speculative negative event of the *cause* class, *increase* a positive polarity version of *associate/correlate*.

We have also given an overview of our current approach to the extraction of interactions between pharmacogenomical entities like drugs, genes and diseases. Our approach is based on syntactic analysis. We have used the top words of the syntactic classes to explore the patterns of domain events, and we suggest a set of interaction labels for the pharmacogenomics domain. Future research will show whether they can be detected reasonably well by text mining approaches, as we speculate.

5. Acknowledgements

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland.

6. References

John Carroll, Guido Minnen, and Edward Briscoe. 2003. Parser evaluation: using a grammatical relation annotation scheme. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 299–316. Kluwer, Dordrecht.

Andrew B Clegg and Adrian J Shepherd. 2007. Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics*, 8:24.

K. Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, John Pestian, Jun'ichi Tsujii, and Bonnie Webber, editors. 2009. *Proceedings of the BioNLP 2009 Workshop*. Association for Computational Linguistics, Boulder, Colorado, June.

G. Erkan, A. Ozgur, and D. R. Radev. 2007. Extracting interacting protein pairs and evidence sentences by using dependency parsing and machine learning techniques. In *Proceedings of BioCreAtIvE 2*.

K. Fundel, R. Küffner, and R. Zimmer. 2007a. RelEx – relation extraction extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007b. RelEx — Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of EACL 2006*.

C. Goddard. 1998. Bad arguments against semantic primitives. *Theoretical Linguistics*, 24:129–156.

Katri Haverinen, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. 2008. Accurate conversion of dependency parses: targeting the Stanford scheme. In *Proceedings of Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, Turku, Finland.

Kaarel Kaljurand, Gerold Schneider, and Fabio Rinaldi. 2009. UZurich in the BioNLP 2009 Shared Task. In *Proceedings of the BioNLP workshop, Boulder, Colorado*.

S. Kim, J. Yoon, and J. Yang. 2008. Kernel approaches for genic interaction extraction. *Bioinformatics*, 9:10.

Adam Meyers, Catherine Macleod, Roman Yangarber, Ralph Grishman, Leslie Barrett, and Ruth Reeves. 1998. Using NOMLEX to produce nominalization patterns for information extraction. In *Coling-ACL98 workshop Proceedings: the Computational Treatment of Nominals*, Montreal, Canada.

George A Miller, R Beckwidth, Ch Fellbaum, D Gross, and K Miller. 1990. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):234–244. private Kopie.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50.

D. Rebholz-Schuhmann, H.Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P.Stoehr. 2006. EBIMed – text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23(2):e237 – e244.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. 2006. An environment for relation mining over richly annotated corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, Christos Andronis, Ourania Konstandi, and Andreas Persidis. 2007. Mining of Functional Relations between Genes and Proteins over Biomedical Scientific Literature using a Deep-Linguistic Approach. *Journal of Artificial Intelligence in Medicine*, 39:127–136.

Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. 2008. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13.

Fabio Rinaldi, Simon Clematide, and Gerold Schneider. 2010a. Ontogene in calbc. In *Proceedings of the CALBC workshop.*

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, and Martin Romacker. 2010b. OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):472–480.

Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. 2012. Using ODIN for a PharmGKB re-validation experiment. *Database: The Journal of Biological Databases and Curation*. accepted for publication.

Katrin Sangkuhl, Dorit S. Berlin, Russ B. Altman, and Teri E. Klein. 2008. PharmGKB: Understanding the effects of individual genetic variants. *Drug Metabolism Reviews*, 40(4):539–551. PMID: 18949600.

Farzaneh Sarafraz and Goran Nenadic. 2010. Using svms with the command relation features to identify negated events in biomedical literature. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 78–85, Uppsala, Sweden, July. University of Antwerp.

Gerold Schneider, Kaarel Kaljurand, and Fabio Rinaldi. 2009. Detecting Protein/Protein Interactions using a parser and linguistic resources. In *CICLing 2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.

Gerold Schneider, Simon Clematide, and Fabio Rinaldi. 2011. Detection of interaction articles and experimental methods in biomedical literature. *BMC Bioinformatics*, 12(Suppl 8):S13.

Gerold Schneider. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Sofie Van Landeghem, Yvan Saeys, Bernard De Baets, and Yves Van de Peer. 2008. Extracting protein-protein interactions from text using rich feature vectors and feature selection. In Tapio Salakoski, Dietrich Rebholz-Schuhmann, and Sampo Pyysalo, editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008),*

Turku, Finland, pages 77–84. Turku Centre for Computer Science (TUCS).

A. Wierzbicka. 1996. *Semantics: Primes and Universals*. Oxford University Press.